LA-UR-19-22194

Approved for public release; distribution is unlimited.

Title: User Behaviour Analytics

Author(s): Turcotte, Melissa
Sanna Passino, Francesco
Moore, Juston Shane
Heard, Nicholas

Intended for: Client presentation

Issued: 2019-03-12

Disclaimer:
Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by Triad National Security, LLC for the National Nuclear Security Administration of U.S. Department of Energy under contract 89233218CNA000001. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

User Behaviour Analytics

Melissa Turcotte
Francesco Sanna Passino, Juston Moore, Nicholas Heard

Advanced Research in Cyber Systems
Los Alamos National Laboratory
LA-UR

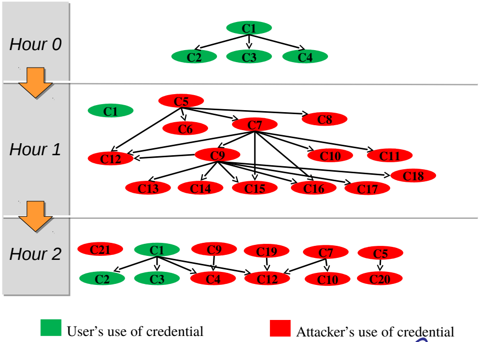
Los Alamos
NATIONAL LABORATORY

User Behaviour Analytics

User Behaviour Analytics is the tracking, collecting and assessing of user data and activities.

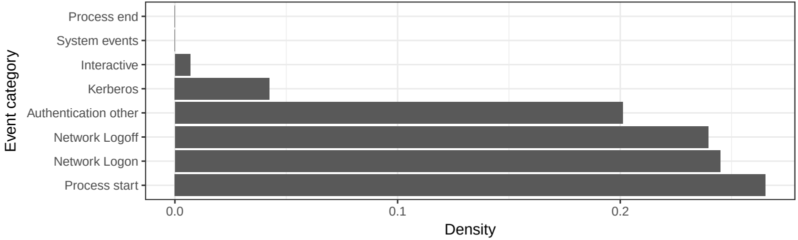
Goal: Detect misuse of user credentials.

- Develop probability models for normal user credential behaviour based on their historical and current network usage.
- Use these models to detect anomalous departures from normal behaviour.



Los Alamos NATIONAL LABORATORY

Data Source - Computer Event Logs



Computer event logs are a critical resource for investigating security incidents.

- authentication, logons
- processes
- applications/services

Many of these log entries are tied to a user credential action.

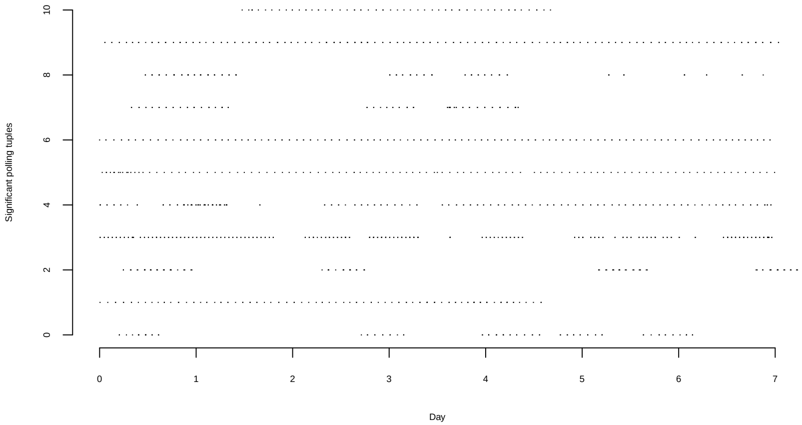
Data considerations

Events removed from analysis:

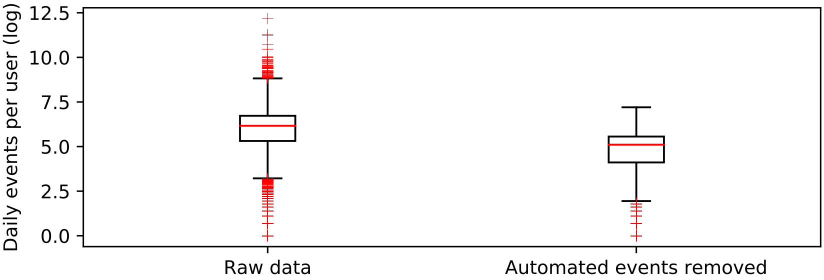
- Account Logoffs;
- Duplicate events/occurring within 30 seconds of each other;
- Events where the originating client is masked by an intermediate e.g. VPN, terminal server, proxy;
- Automated events.

Filtering automated events

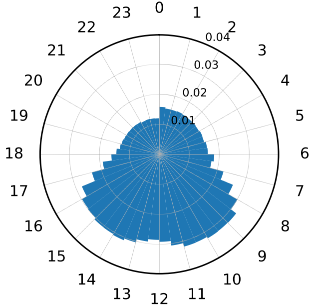
Each unique tuple (user, client, server, eventid) has an associated time series, want to remove those that have strong periodic behaviour.

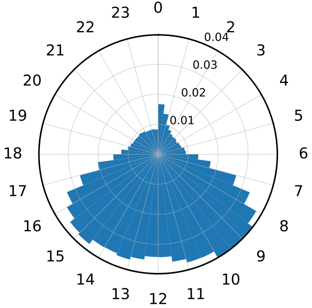


- Calculate the discrete Fourier transform (DFT) for the counting process associated with each time series using the fast Fourier transform.
- Use Fisher's g-test to determine significance of maximum peak in the DFT.
- Get an estimate for the most common polling frequency.
- Remove all events that occur within that frequency.



Boxplot of number of events per day for each user.





Distribution of daily activity of all user events estimated to 30 minute bins.

Los Alamos
NATIONAL LABORATORY

Future Data Sources

- Badge reader data
- HR data
- proxy logs
- e-mail logs

Los Alamos
NATIONAL LABORATORY

User Behaviour anomaly detection

Initial modelling effort only uses the following event fields for analysis:

- client computer, denoted X;
- server computer, denoted Y;
- event type, denoted E.

Extra fields relating to some event types can be added to the model, such as if the event is a process, what executable.

We model the sequence of events for each user. Modelling the precise times of the events is the subject of current research - paper forthcoming.

Model

For each user credential a sequence of events is observed over time:

$$\{(X_t, Y_t, E_t) : t = 1, 2, \ldots\}$$

- X_t = client, $X_t \in V$,
- Y_t = server, $Y_t \in V$,
- E_t = event type, $E_t \in E$,

where V is the set of computers in the network and E the set of possible event types.

Want to obtain a p-value (anomaly score) for each observed (X_t, Y_t, E_t).

Split up the model into two components:

- "New" behaviour - user using a computer never used before;
  - Degree based popularity model.
- User uses a computer that it's used before.
  - Since the above variables are all categorical, these are most simply modelled using Bayesian techniques for the category probabilities. Provides a flexible framework, with simple updating.

For each new observed event, use the probability models to obtain a score for how likely the observed event is according to the users historical behaviour.

Los Alamos
NATIONAL LABORATORY

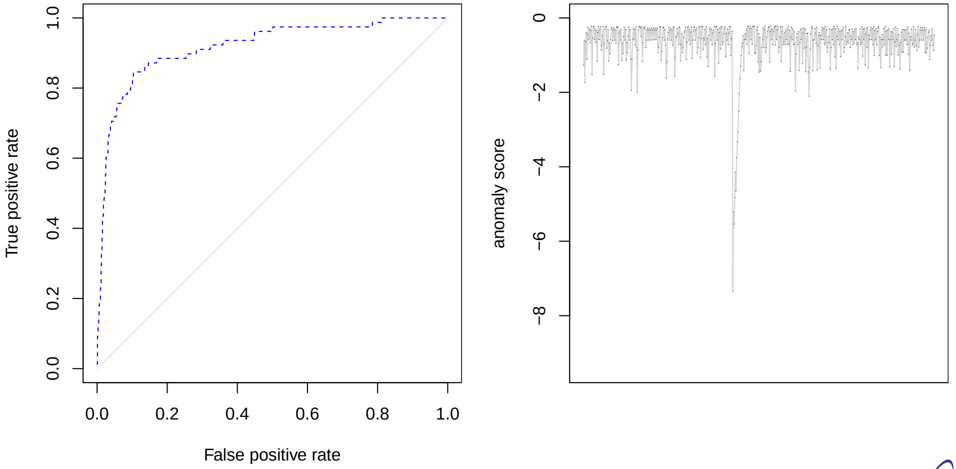
Red team data

- Two months of data: 443,934,073 events for 10,759 user credentials.
- Month long red-team exercise in the second month of data, 78 known compromised credentials.
- Random selection of 1,000 credentials used to demonstrate the method, plus compromised credentials $\rightarrow$ 50,536,677 associated events.

Data available at `http://csr.lanl.gov/data/cyber1/`

Detection of the red team attack

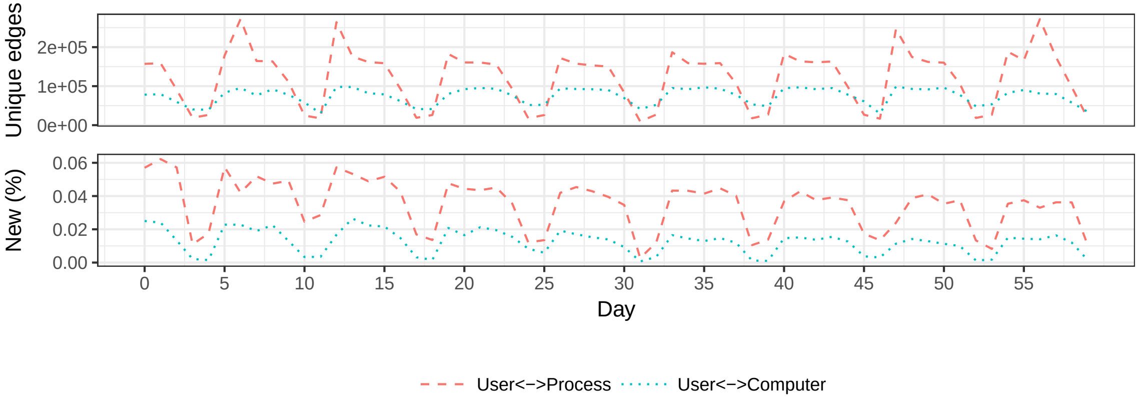
ROC curve and anomaly scores over time for a compromised user.



Los Alamos NATIONAL LABORATORY

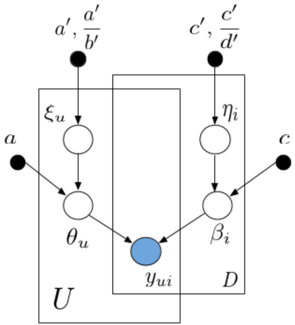
Peer-based anomaly detection

- "New" user behaviour currently modelled using node-based popularity - can we do better?
- Incorporating peer-based analysis allows anomaly detection systems to leverage the behaviour of similar users (peers) to better identify individual profiles, reducing false alarms.



Recommender systems

- Utilise recommender system algorithms to predict user actions that are unlikely based on peer-group preferences.
- Allows for different peer groups depending what features of the data are being considered.



The hierarchical Poisson factorization model

- For n users and m items, let $\boldsymbol{Y}$ be a $n \times m$ matrix of counts where each element Y_{ui} is the random variable for the number of times user u interfaced with item i.
- k-dimensional Poisson factorization model:
  - User u represented by R latent factors $\theta_u = (\theta_{u1}, \ldots, \theta_{uR})$,
  - Item i represented by R latent factors $\beta_i = (\beta_{i1}, \ldots, \beta_{iR})$,

$$Y_{ui} \sim \text{Poisson}\left(\sum_{r=1}^{R} \theta_{ur}\beta_{ir}\right).$$

Covariates

Node labels are often available, want to use this information to improve predictive performance.

Extend the model to include covariate information.

$$Y_{ui} \sim \text{Poisson}\left(\sum_{r=1}^{R} \theta_{ur}\beta_{ir} + \sum_{k=1}^{K}\sum_{h=1}^{H} \phi_{kh} x_{uk} y_{ih} \right)$$

The additional parameters in the model can be interpreted as follows:

- $\phi_{kh} \in \mathbb{R}_+$ is an interaction between the k-th user covariate and the h-th host/item covariate,
- $x_{uk} \in \{0, 1\}$ and $y_{ih} \in \{0, 1\}$ are binary variables denoting whether the user (or item) possesses covariate k (or h).

Los Alamos
NATIONAL LABORATORY

Data

Two data sets:

- Users and the computers they authenticate *from* - `userSource`,
- Users and the computers they authenticate *to* - `userDest`.

Features:

- Approximately $12,000$ users and $15,000$ computers;
- Split the data into a training set and test set;
- $\approx 15\%$ "new" edges for userDest and $\approx 25\%$ for userSource in test set.

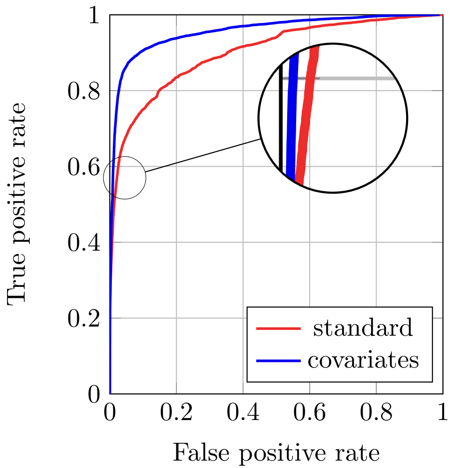
Covariates:

- User Covariates - Credential type, Group, Job title, Location
- Computer Covariates - Group of owner, Subnet, Type

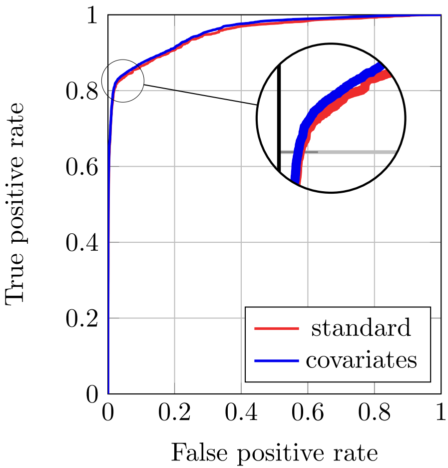
Data available at `http://csr.lanl.gov/data/2017.html`

Los Alamos NATIONAL LABORATORY

ROC curves for the standard model and the extended model with covariates.

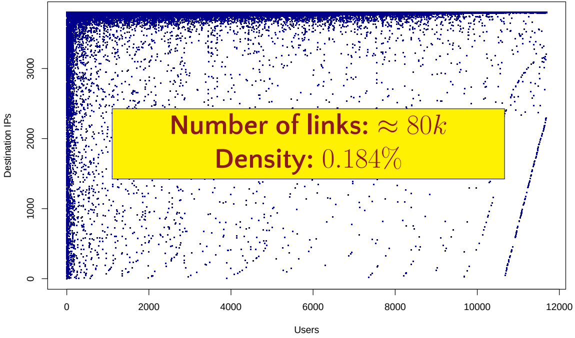


userDest

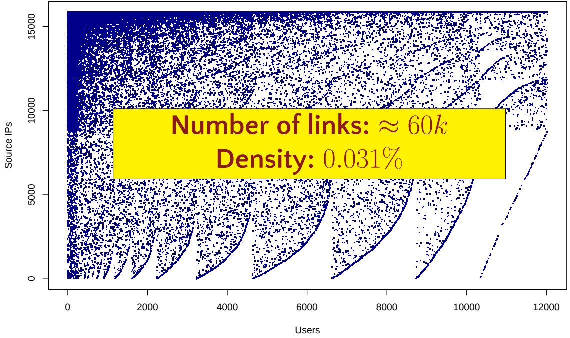


userSource

Adjacency matrix



`userDest`



`userSource`

Cold Starts

Cold-start: observe a new user or host in the test set, never seen before $\rightarrow$ how can we provide reliable estimates?

- If we know the covariates for the new user or item, can use the learned covariate parameters to make predictions.

$\approx 92\%$ prediction accuracy for both `userDest` and `userSource`.

Los Alamos
NATIONAL LABORATORY

Path Forward

- Combine the two approaches above to provide a robust overall model for UBA.
- Utilise more data sources to get a more holistic view.